VISION FOR BLINDS

REVIEW - 3

Course Code: CSE4020-Machine Learning

Slot: B1+TB1

Prof: Vijayarajan V

Sarthak Surana 17BCE0996

Dhruv Mittal 17BCE2110

Abstract

We see the things through our eyes. We understand the things we see continuously. This process is so easy and normal for us that we don't even think about it. Understanding our surroundings is very easy task for a normal human. We are blessed with eyes which help us perceive this beautiful world. But what about those who are not as lucky as we are. The Blinds. They have to completely rely on their other senses to get a basic understanding of the world. They 'see' the world very differently. Life is a lot tougher for them. They cannot enjoy the beauty of nature. So, we propose to develop an algorithm through which we can help them 'see'. Understanding an image is not very easy for a machine. Understanding the content of image and translating it into natural language is very challenging and computationally expensive. Recently, deep learning methods have displaced classical methods and are achieving astonishing results in generation of automatic description of image. We propose to use this automated image description generator to be used to recite what the image means. With advancement in technology, introduction of quantum computers or some other technological advancements, such image description generation could be done in real time. This will eventually be able to generate captions for what the system sees. This kind of technology has the potential to change the world for blinds and help them lead a normal life.

In this project we create an automated image captioning model using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to produce a series of texts that best describe the image. This model is trained on Flickr 8k dataset. Image captioning requires that you create a complex deep learning model with two components: a CNN that transforms an input image into a set of features, and an RNN that turns those features into rich, descriptive language. This project uses the encoder-decoder model, in which the CNN which performs the feature extraction and the output of the encoder is fed to the decoder which processes the classified features into appropriate sentences. The feature extraction will be done by the latest Inception V3 module-50 technology with means of transfer learning so that we can modify the project specific to our purpose. The language model uses natural language toolkit for simple natural language processing and the architecture used for recurrent neural network is long-short term memory.

Keywords:

Image captioning, CNN, RNN, encoder - decoder, Deep learning, LSTM, Flickr 8k dataset.

Introduction

In our project we aim to create an auto captioning model for an image to create vision and audio for blind people. We achieve this by using Convolutional Neural Network for image feature classification and object localization. Then we use Recurrent Neural Networks for text generation. We use Long Short-Term Memory (LSTM) a type of RNN model which can learn from sequential data like a series of words and characters. These networks use hidden layers that link the input and output layers. Which creates a memory loop so that the model can learn from previous outputs. So basically, CNN works with spatial features and RNN helps in solving sequential data.

This project uses advanced methods of computer vision using Deep Learning and natural language processing using a Recurrent Neural Network. Deep Learning is a machine learning technique with which we can program the computer to learn complex models and understand patterns in a large dataset.
The combination of increasing computation speed, wealth of research and the rapid growth of technology. Deep Learning and AI is experiencing massive growth worldwide and will perhaps be one of the world's biggest industries in the near future. The 21st century is the birth of AI revolution, and data becoming the new 'oil' for it. Every second in today's world large amounts of data is being generated. We need to build models that can study these datasets and come up with patterns or find solution for analysis and research. This can be achieved solely due to deep learning.

Computer Vision is a cutting-edge field of computer science that aims to enable computers to understand what is being seen in an image. Computers don't perceive the world like humans do. For them the perception is just sets of raw numbers and because of several limitations like type of camera, lighting conditions, clarity, scaling, viewpoint variation etc. make computer vision so hard to process as it is very tough to build a robust model that can work on every condition.

The neural network architectures normally we see were trained using the current inputs only. We did not consider previous inputs when generating the current output. In other words, our systems did not have
any memory elements. RNNs address this very basic and important issue by using memory (i.e. past inputs to the network) when producing the current output. These are the things we will be learning ahead in this project.

Literature Review

1. **A Survey on Automatic Image Caption Generation by:**

   Shuang Bai∗

Image captioning approach robotically producing a caption for a photo. As a lately emerged research place, its miles attracting more and more interest. To achieve the purpose of picture captioning, semantic records of pictures desires to be captured and expressed in natural languages. Connecting both studies communities of computer vision and herbal language processing, image captioning is a pretty difficult project. Various tactics have been proposed to remedy this trouble. In this paper, we gift a survey on advances in picture captioning research. Based on the method followed, we classify image captioning procedures into different classes. Representative strategies in each category are summarized, and their strengths and limitations are mentioned. In this paper, we first talk techniques utilized in early paintings which are in particular retrieval and template based totally. Then, we consciousness our essential attention on neural network primarily based techniques, which provide state of the artwork consequences. Neural community-based methods are further divided into subcategories primarily based on the specific framework they use. Each subcategory of neural community based totally methods are mentioned in element. After that, state of the art strategies is as compared on benchmark datasets. Following that, discussions on destiny research instructions are provided

2. **Image Captioning with Convolutional Neural Networks**

   By: Michal Najman

In this thesis, we difficult on picture captioning concerning specifically dense image captioning. We gift technical basics of a model striving to clear up any such challenge. Concretely, an in-depth shape of Dense Cap and Neural Image Caption is discussed. Experimentally, we have a look at effects of Dense Cap and analyses the model's weaknesses. We display that 92% of the generated captions are same to a caption within the education set even as the best of those and the radical ones remains the same. We propose a criterion that significantly reduces a fixed of captions addressing an image whilst SPICE rating of the set is maintained.

3. Improving Image Captioning by Leveraging Knowledge Graphs

By: Yimin Zhou, Yiwei Sun, Vasant Honavar Artificial Intelligent Research Laboratory the Pennsylvania State University

We explore the use of a knowledge graphs that capture general or commonsense knowledge, to augment the information extracted from images by the state-of-the-art methods for image captioning. The results of our experiments, on several benchmark data sets such as MS COCO, as measured by CIDEr-D, a performance metric for image captioning, show that the variants of the state-of-the-art methods for image captioning that make use of the information extracted from knowledge graphs can substantially outperform those that rely solely on the information extracted from images.

CNet-NIC uses YOLO9000, a state-of-the-art general-purpose real-time object recognition module that is trained to recognize 9000 object categories. YOLO9000 takes an image as input and produces as output, a collection of terms that refer to objects in the scene. CNet-NIC use an external knowledge graph, specifically, Concept Net a labeled graph which connects words and phrases of natural language connected by edges that denote commonsense relationships between them, to infer two sets of terms related to the words that describe the objects found in the scene by the object recognition module. The first set of terms are retrieved based on the individual objects in the scene. The second set of terms are retrieved based on the entire collection of objects in the scene. The resulting terms are then provided to a pre-trained RNN to obtain the corresponding vector space embedding of the terms. A CNN is used to obtain vector space embedding of the image features. The two resulting vector space embeddings are used to specify the initial state of an LSTM-based RNN which is trained to produce the caption for the input image.

4. Image Captioning with Object Detection and Localization

By: Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Department of Electronic Engineering, Tsinghua University, Beijing

Automatically generating a natural language description of an image is a task close to the heart of image understanding. In this paper, we present a multi-model neural network method closely related to the human visual system that automatically learns to describe the content of images. Our model consists of two sub-models: an object detection and localization model, which extract the information of objects and their spatial relationship in images respectively;

Besides, a deep recurrent neural network (RNN) based on long short-term memory (LSTM) units with attention mechanism for sentences generation. Each word of the description will be automatically aligned to different objects of the input image when it is generated. This is similar to the attention mechanism of the human visual system. Experimental results on the Flickr 8k dataset showcase the merit of the proposed method, which outperform previous benchmark models.

5. Convolutional Image Captioning

By: Jyoti Aneja, Aditya Deshpande, Alexander G. Schwing
University of Illinois at Urbana-Champaign

In recent years significant progress has been made in image captioning, using Recurrent Neural Networks powered by long-short-term-memory (LSTM) units. Despite mitigating the vanishing gradient problem, and despite their compelling ability to memorize dependencies, LSTM units are complex and inherently sequential across time. However, the complex addressing and overwriting mechanism combined with inherently sequential processing, and significant storage required due to back-propagation through time (BPTT), poses challenges during training. Despite the fact that the above RNNs based on LSTM/GRU deliver remarkable results, e.g., for image captioning, their training procedure is all but trivial. For instance, while the forward pass during training can be in parallel across samples, it is inherently sequential in time, limiting the parallelism. To address this issue, they proposed a Pixel CNN architecture for conditional image generation that approximates an RNN. This article demonstrates that convolutional architectures with attention achieve state-of
the-art performance on machine translation tasks. In spirit similar is our approach for image captioning, which is convolutional but addresses a different task.

6. Graph-based Automatic Image Captioning

By:Jia-Yu Pan Carnegie Mellon University, Pittsburgh, PA

Given an image, how do we automatically assign keywords to it? In this paper, we propose a novel, graph-based approach (GCap) which outperforms previously reported methods for automatic image captioning. Moreover, it is fast and scales well, with its training and testing time linear to the data set size. We report auto-captioning experiments on the “standard” Corel image database of 680 MBytes, where GCap outperforms recent, successful auto-captioning methods by up to 10 percentage points in captioning accuracy.

7. **Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning**

By: Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut ,Google AI

We present a new dataset of image caption annotations, Conceptual Captions, which contains an order of magnitude more images than the MS-COCO dataset (Lin et al., 2014) and represents a wider variety of both images and image caption styles. We achieve this by extracting and filtering image caption annotations from billions of webpages. We also present quantitative evaluations of a number of image captioning models and show that a model architecture based on Inception-ResNetv2 (Szegedy et al., 2016) for image-feature extraction and Transformer (Vaswani et al., 2017) for sequence modeling achieves the best performance when trained on the Conceptual Captions dataset.

8. **Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering**

By: Peter Anderson, Xiaodong He

Top-down visual attention mechanisms have been used extensively in image captioning and visual question answering (VQA) to enable deeper image understanding through fine-grained analysis and even multiple steps of reasoning. In this work, we propose a combined bottom-up and topdown attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. This is the natural basis for attention to be considered. Within our approach, the bottom-up mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. Applying this approach to image captioning, our results on the MSCOCO test server establish a new state-ofthe-art for the task, achieving CIDEr / SPICE / BLEU-4 scores of 117.9, 21.5 and 36.9, respectively. Demonstrating the broad applicability of the method, applying the same approach to VQA we obtain first place in the 2017 VQA Challenge

9. **Automatic image captioning**

By: Jia-Yu Pan Dept. of Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA Hyung-Jeong Yang Dept. of Comput. Sci., Carnegie Mellon Univ.,

Pittsburgh, PA, USA P. Duygulu C. Faloutsos

We examine the problem of automatic image captioning. Given a training set of captioned images, we want to discover correlations between image features and keywords, so that we can automatically find good keywords for a new image. We experiment thoroughly with multiple design alternatives on large datasets of various content styles, and our proposed methods achieve up to a 45 percent accuracy.

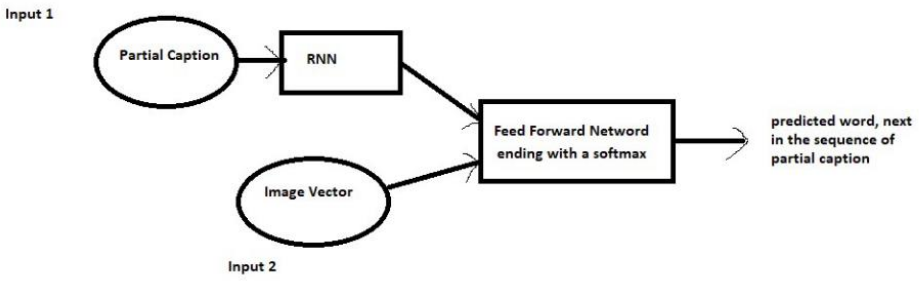
10. Auto-Encoding Scene Graphs for Image Captioning
By: Xu Yang, Kaihua Tang, Hanwang Zhang, Jianfei Cai;

The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10685-10694 We propose Scene Graph Auto-Encoder (SGAE) that incorporates the language inductive bias into the encoder-decoder image captioning framework for more human-like captions. Intuitively, we humans use the inductive bias to compose collocations and contextual inference in discourse. For example, when we see the relation ”person on bike”, it is natural to replace ”on” with ”ride” and infer ”person riding bike on a road” even the ”road” is not evident. Therefore, exploiting such bias as a language prior is expected to help the conventional encoder-decoder models less likely to overfit to the dataset bias and focus on 5 reasoning. Specifically, we use the scene graph — a directed graph (G) where an object node is connected by adjective nodes and relationship nodes — to represent the complex structural layout of both image (I) and sentence (S). In the textual domain, we use SGAE to learn a dictionary (D) that helps to reconstruct sentences in the pipeline, where D encodes the desired language prior; in the vision-language domain, we use the shared D to guide the encoderdecoder in the pipeline. Thanks to the scene graph representation and shared dictionary, the inductive bias is transferred across domains in principle. We validate the effectiveness of SGAE on the challenging MS-COCO image captioning benchmark, e.g., our SGAE-based single-model achieves a new state-of-the-art 127.8 CIDEr-D on the Karpathy split, and a competitive 125.5 CIDEr-D (c40) on the official server even compared to other ensemble models. Code has been made available at: https://github.com/yangxuntu/SGAE

Proposed work and methodology /Architure Model

Working

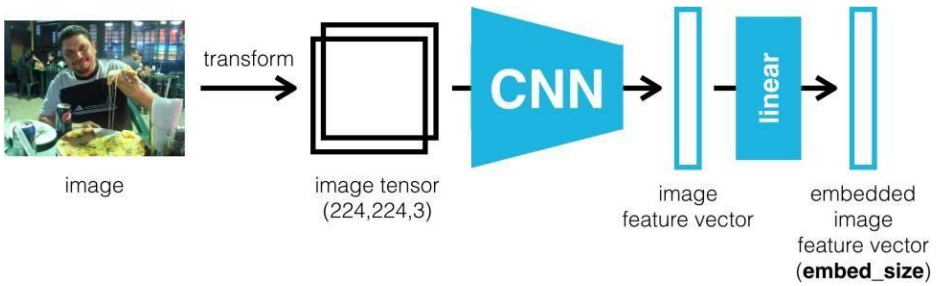
Image captioning requires that you create a complex deep learning model with two components: a CNN that transforms an input image into a set of features, and an RNN that turns those features into rich, descriptive language and audio, hence we convert the captions into audio.



High level architecture

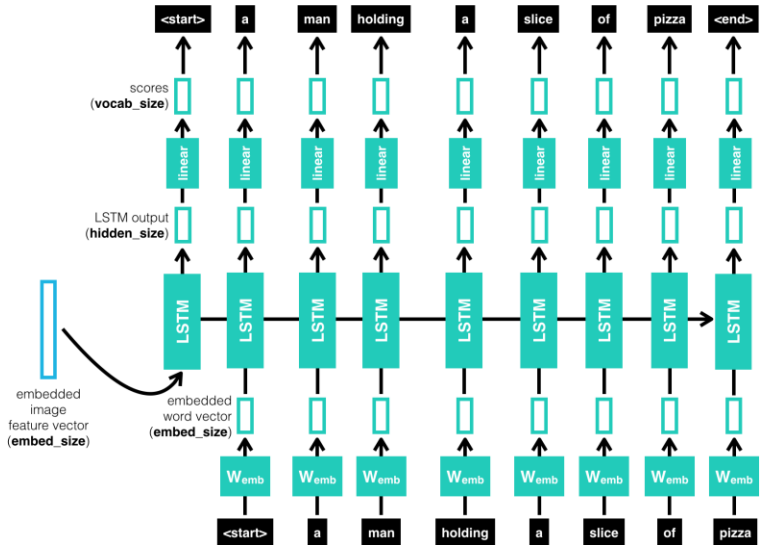
CNN Encoder

The encoder is based on a Convolutional neural network that encodes an image into a compact representation. The CNN-Encoder is an Inception V3 module (Residual Network). These kinds of network help regarding to the vanishing and exploding gradient type of problems. The main idea relies on the use of skip connections which allows to take the activations from one layer and suddenly feed it to another layer, even much deeper in the neural network and using that, we can build Inception V3 modules which enables to train very deep networks.



RNN Decoder

The CNN encoder is followed by a recurrent neural network that generates a corresponding sentence. The RNN-Decoder consists in a single LSTM layer followed by one fully-connected (linear) layer. The RNN network is trained on the Flickr 8k dataset. It is used to predict the next word of a sentence based on previous words. The captions are presented as a list of tokenized words so that the RNN model can train and back propagate to reduce errors and generate better and more understandable texts describing the image.



Flickr 8k dataset

Flilckr8K contains 8,000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events. The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.

Inception V3 module

There are 4 versions. The first Google Net must be the Inception-v1 , but there are numerous typos in Inception-v3 which lead to wrong descriptions about Inception versions. These maybe due to the intense ILSVRC competition at that moment. Consequently, there are many reviews in the internet mixing up between v2 and v3. Some of the reviews even think that v2 and v3 are the same with only some minor different settings.

Hardware And Software Requirements

Hardware

- Processor: Minimum 1 GHz; Recommended 2GHz or more.
- Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
- Hard Drive: Minimum 32 GB; Recommended 64 GB or more.
- Memory (RAM): Minimum 1 GB; Recommended 4 GB or above.

Software

- Python
- OpenCV
- Keras
- TensorFlow
- Atom editor
- Windows/Linux

Dataset / Tools Used

Flickr 8k dataset which comes with images and captions for supervised learning. Every image contains multiple caption which all are relevant to the picture. Our dataset used here is 8k. With more computational power we can go for 30k dataset which is 150 times bigger than our dataset. This dataset is only used for demonstration purpose.

Implementation and Experiment Results:

In video recording the recorder did not capture the voice of auto captioning audio but it's there and working perfectly fine. The accuracy achieved in image captioning is 77%. Which is very decent.

```
In [72]: speaker = win32com.client.Dispatch("SAPI.SpVoice")

In [92]: for i in range(6):
             rn =  np.random.randint(0, 1000)
             img_name = list(encoding_test.keys())[rn]
             photo = encoding_test[img_name].reshape((1,2048))

             i = plt.imread(images+img_name)
             plt.imshow(i)
             plt.axis("off")
             plt.show()

             caption = predict_caption(photo)
             speaker.Speak(caption)
             print(caption)
```

basketball player in white is being thrown by the ball

black dog is running on the grass

basketball player in white is shooting by the ball

man rides motorcycle on dirt track

group of people are standing in front of some

group of people are standing in front of some

black and white dog is running on the grass

Future Scope:

This technique can be used to make software installed on spectacles for visually impaired people. So, the spectacle would take the sight as its input and produce audio as captions so that the person wearing would know about the surrounding.

Conclusion:

We combine "Image Labeling" and "Automatic Machine Translation" into an end-to-end hybrid neural network system. The developed model is capable to autonomously view an image and generate a reasonable description in natural language with reasonable accuracy and naturalness. Further extension of the present model can be in regard to increasing additional CNN layers or increasing/implementing pre-training, which could improve the accuracy of the predictions.

References:

[1] P. Anderson, Fernando, Johnson, and S. Gould. Spice:Semantic propositional image caption evaluation. In ECCV, 2016

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In International Conference on Computer Vision (ICCV), 2015

[3] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014

[4] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Int. Res., 47(1), May 2013.

[4] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (mrnn). CoRR, abs/1412.6632, 2014.

[5] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In 2015 IEEEConference on Computer Vision and Pattern Recognition (CVPR), pages 3376–3385, June 2015.

[6] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13, pages III–1310–III–1318. JMLR.org, 2013.